StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis

Abstract

Text-to-image synthesis has recently seen significant progress thanks to large pretrained language models, large-scale training data, and the introduction of scalable model families such as diffusion and autoregressive models. However, the best-performing models require iterative evaluation to generate a single sample. In contrast, generative adversarial networks (GANs) only need a single forward pass. They are thus much faster, but they currently remain far behind the state-of-the-art in large-scale text-to-image synthesis. This paper aims to identify the necessary steps to regain competitiveness. Our proposed model, StyleGAN-T, addresses the specific requirements of large-scale text-to-image synthesis, such as large capacity, stable training on diverse datasets, strong text alignment, and controllable variation vs. text alignment tradeoff. StyleGAN-T significantly improves over previous GANs and outperforms distilled diffusion models - the previous state-of-the-art in fast text-to-image synthesis - in terms of sample quality and speed.